

プロジェクト名(タイトル): 深層学習などによる NMR データおよび生体高分子の解析

利用者氏名: 小林直宏

理研における所属研究室名: 生命機能科学研究センター先端 NMR 開発・応用研究チーム

1. 本プロジェクトの研究の背景、目的、関係するプロジェクトとの関係

NMR 解析法は他の構造解析法として X 線結晶解析やクライオ電顕解析などに比べて自動解析法は大きく遅れを取っている。HOKUSAI のような大規模 PC クラスタの利用は、NMR のデータ解析の自動化および高度化、深層学習ツールの開発などへの応用に不可欠であり、特に Amber20 による MD 計算は高速化が期待できる。現在では MD 計算で得られる生体高分子の構造座標と NMR 解析によって得られる化学シフトデータの関連性を機械学習させることを目的とし、2023 年度は MD 計算を行うことを重点に置いた。

研究代表者は HOKUSAI に搭載されている 40~200 Intel-Core を利用することで大規模な MD 並列計算を実行し、計算結果を深層学習の教師ありデータとして多数実行した。

2. 具体的な利用内容、計算方法

Amber20 による explicit water MD 計算 (20-40 core), を 8~20 ノード用いて、温度レプリカ交換法 (T-REMD), および 自身で開発した構造から化学シフトを予測する深層学習器を実行利用した。HOKUSAI には GPU が搭載されていないため、深層学習訓練計算は自身で構築した GPU workstation で実行し、大規模ストレージとコア数を必要とする計算は HOKUSAI によって手分けして実行させた。

3. 結果

構造から化学シフト予測する方法論は以前より公開されており、既存のツールでも浅めの機械学習を応用している SPARTA+, UCFS-Shift と呼ばれるツールにおいてもある程度の予測精度は実現されている。代表者により開発された新規深層学習器によるベンチマークは結果としては UCFS-Shift と同等かやや優れている程度であったため、更なる改良を行うことで2段型の化学シフト予測ツールのプロトタイプを完成した。一方で、局所運動平均による運動性が予測精度を下げている事が徐々に明らかになってきた。これを克服すべく、運動性が NMR 実験により明らかとなっている複数の小型のタンパク質について長時間 MD 計算を行った。これらは細かい時定数での trajectory が得られており、

今年度中に運動性時定数解析を行う予定である。

4. まとめ

代表者により開発された深層学習器による構造座標からの化学シフト予測性能は既存のツール SPARTA+, UCFS-Shift との大きな差別化には成功していなかったが、今年度の改良により2世代目の予測器として完成したため関連アルゴリズムに関する論文を鋭意執筆中である。

また、運動性による効果は更なる興味深い展開を期待できるため、次年度に向けた3世代目の深層学習器の設計を行っている。

5. 今後の計画・展望

代表者により開発された深層学習器は、今年度の改良によって性能向上は実現したが、実用性には疑問が残る。運動性解析による時定数を考慮した3世代目の深層学習器を設計し、大規模な MD 計算の実行を次年度以降も継続することで、より高性能な予測性能獲得が期待できる。また、逆問題にも取り組むことで MD 計算や NMR による運動性解析を実行することなく、深層学習のみで高い分解能の運動性を予測するという新規深層学習システム的设计開発が実現すれば、革新的な第4世代システムとして完成できることが期待できる。

6. 利用がなかった場合の理由